VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ
DEPARTMENT OF COMPUTER SYSTEMS

# BIOINFORMATICKÝ NÁSTROJ PRO PREDIKCI ROZPUSTNOSTI PROTEINŮ

BIOINFORMATICS TOOL FOR PREDICTION OF PROTEIN SOLUBILITY

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — Jiří Čermák
AUTHOR

VEDOUCÍ PRÁCE — Ing. Tomáš Martínek, Ph.D.
SUPERVISOR

BRNO 2017

# Abstrakt

Abychom dosáhli levnější a efektivnější výroby proteinů, musíme být schopni predikovat, zda budou proteiny rozpustné. V této práci se zabýváme vytvořením bioinformatických datových sad na základě databází Target Track a eSol, testováním příznaků používaných v existujících nástrojích zabývajících se rozpustností proteinů a tvorbou nového prediktoru. Přestože se nám nedaří vytvořit efektivní nástroj na predikci rozpustnosti proteinů, zjišťujeme, že ve většině případů staré příznaky na nové datové sadě nekorelují s rozpustností proteinů tak silně, jako tomu je u starších a menších datových sad.



# Abstract

To achieve cheaper and more efficient protein production, we must be able to predict protein solubility. In this thesis, we describe creation of bioinformatic data sets based on Target Track and eSol databases, we test the features used in existing protein solubility prediction tools and create a new predictor. Even though we fail to create an effective prediction tool we find out that in most cases the old features tested on the new data do not correlate with protein solubility as strongly as others repot in older and smaller datasets.



# Klíčová slova

Rozpustnost proteinů, Strojové učení, Agregace proteinů, Proteinové inženýrství, Syntéza proteinů

# Keywords

Protein solubility, Machine learning, Protein Aggregation, Protein engineering, Protein synthesis


# Citace


Čermák Jiří: Bioinformatický nástroj pro predikci rozpustnosti proteinů, bakalářská práce, Brno, FIT VUT v Brně, 2017

# Bioinformatický nástroj pro predikci rozpustnosti proteinů

## Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením
Ing. Tomáše Martínka, Ph.D.
Další informace mi poskytl Martin Marušiak
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Jiří Čermák
7.5.2017

## Poděkování

Chtěl bych poděkovat Ing. Tomáši Martínkovi, Ph.D. za vedení práce a pravidelné poskytování konzultací. Dále bych chtěl poděkovat Martinu Marušiakovi za poskytnutí dat.

# Obsah

# 1 Úvod

Rozpustnost proteinů velice úzce souvisí s výrobou funkčních proteinů. Funkční proteiny chceme vyrábět z několika důvodů. Pro vědecké studium proteinů a jejich interakcí je potřeba umět proteiny vyrobit. Pokud nějaký protein vykazuje zajímavé vlastnosti pro průmyslovou aplikaci, je třeba ho vyrábět ve velkém množství.

Existují dva hlavní typy výroby proteinů. Prvním je chemická syntéza, kde se protein přímo syntetizuje. Touto metodou v současné době dokážeme vyrobit pouze malé proteiny, a to pouze v malém množství. Druhou metodou je rekombinantní exprese proteinu, která se provádí buď in vitro (ve zkumavce), nebo in vivo (uvnitř hostitelské buňky). Při výrobě in vitro jsou veškeré molekuly potřebné pro výrobu proteinů extrahovány z buňky a celá syntéza probíhá ve zkumavce. Metoda in vivo využívá již existujících mechanismů exprese genů hostitelského organismu pro výrobu proteinů.

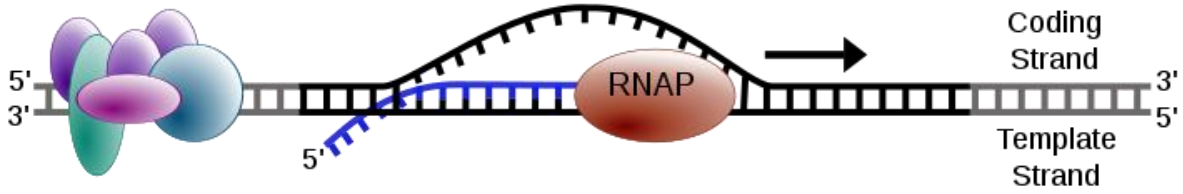
Problém rozpustnosti je tradičně řešen experimentální metodou. Protein se jednoduše vyrobí metodou pokus, omyl. Tento přístup je ale časově a finančně náročný. Je proto výhodné vytvořit nástroj, který umožní již dopředu vyřadit nesolubilní proteiny a zmenšit tak soubor proteinů, jejichž rozpustnost je nutné experimentálně ověřit.

V této práci se budeme zabývat vytvořením nástroje, který bude schopen určit na základě aminokyselinové sekvence proteinu jeho rozpustnost při výrobě in vivo. Nejprve se v teoretické části budeme zabývat metodami výroby proteinů, faktory, které ovlivňují rozpustnost proteinů, existujícími datovými sadami pro rozpustnost proteinů a existujícími nástroji na predikci rozpustnosti proteinů. V praktické části se pak budeme zabývat přímo tvorbou vlastní datové sady, výběrem a testováním příznaků pro strojové učení a konkrétním návrhem prediktoru.

# 2 Teoretická část

## 2.1 Exprese proteinů

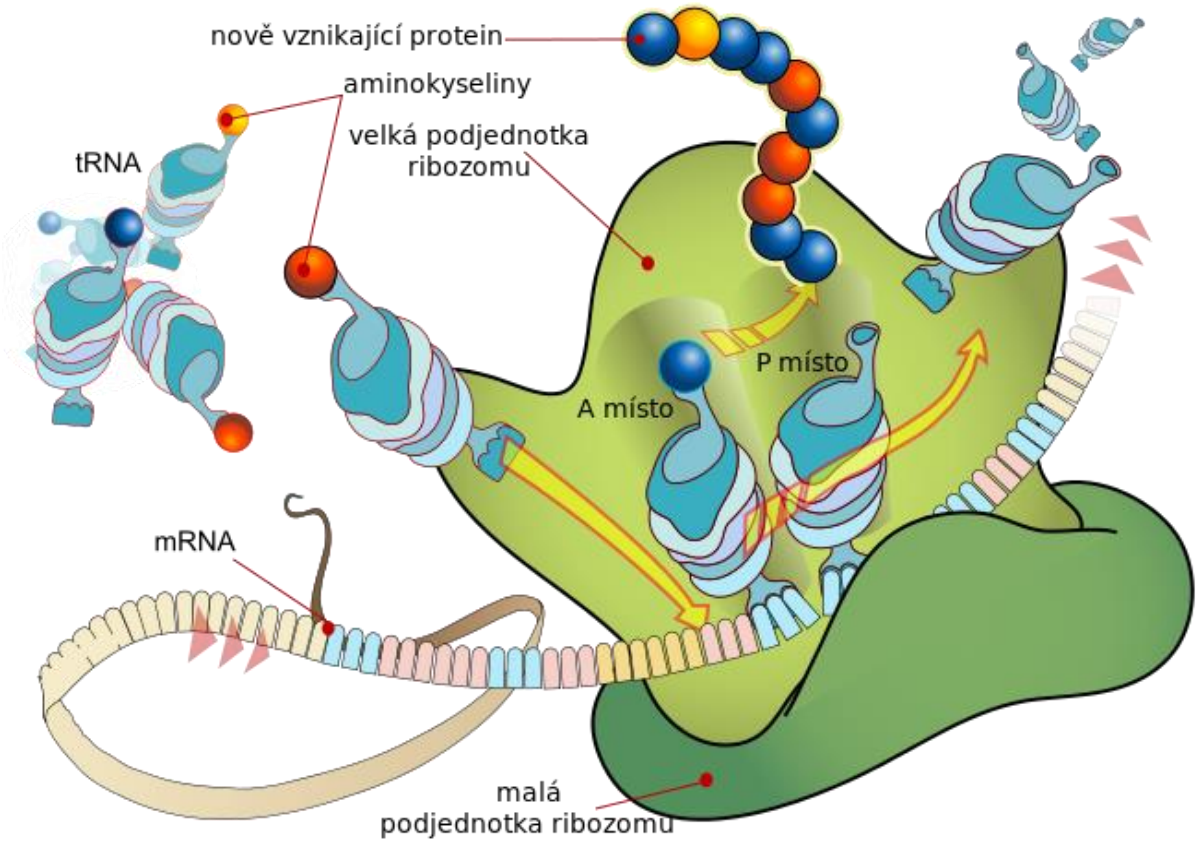
Exprese proteinů je proces, kdy je podle informací v genu syntetizován protein. Exprese je zahájena transkripcí, kdy se enzym nazývaný RNA polymeráza (na obrázku RNAP) naváže na začátek genu. RNA polymeráza poté postupuje postupně po řetězci DNA (na obrázku černě) a přepisuje ho za pomoci dalších molekul v organismu do RNA (na obrázku modře) řetězce (viz obrázek 2-1 - transkripce).




*Obrázek 2-1-Transkripce (https://commons.wikimedia.org/wiki/File:Simple_transcription_elongation1.svg)*

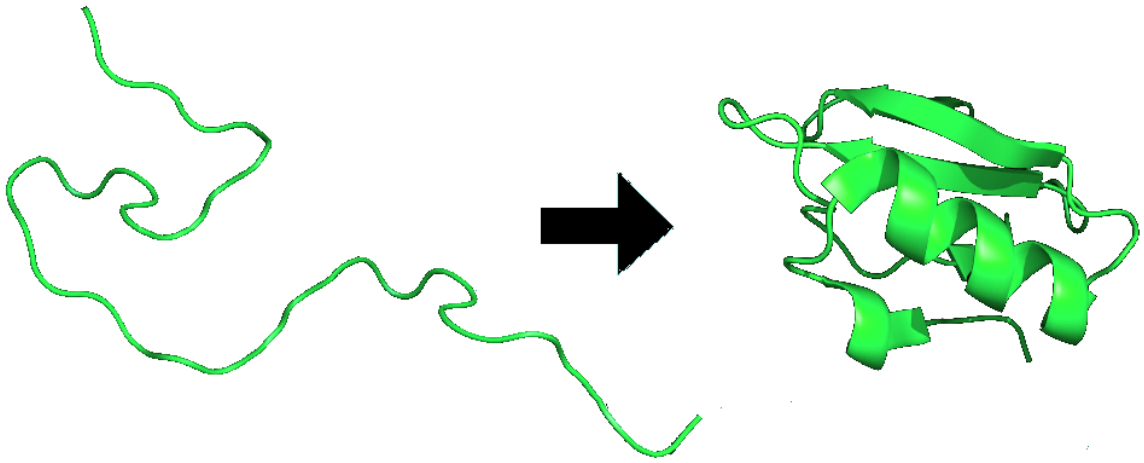
V závislosti na organismu je RNA řetězec poté dále upraven a v eukaryotických buňkách přesunut ven z buněčného jádra do cytoplazmy.

Jakmile je RNA v cytoplazmě následuje proces translace, kdy organela nazývající se ribozom postupně, aminokyselina po aminokyselině vytváří požadovaný protein (viz Obrázek 2-2 - Translace).




*Obrázek 2-2 – Translace (https://commons.wikimedia.org/wiki/File:Ribosome_mRNA_translation_en.svg)*

Ještě během fáze translace započne poslední etapa exprese proteinu a to skládání. Protein vychází z ribozomu ve struktuře lineárního řetězce a pro jeho správné fungování je třeba, aby protein dosáhl nativního stavu (viz obrázek 2-3 - Skládání). Při procesu skládání existuje mnoho faktorů, které ovlivňují, jestli se protein správně složí. Protein v podstatě hledá konformaci, ve kterém je molekula v nejnižším energetickém stavu, a tedy nejstabilnější v daném prostředí. Vliv na úspěšnost skládání proteinu tedy má okolní prostředí, koncentrace solí, pH, teplota a přítomnost dalších proteinů. Velice důležité jsou zejména proteiny nazývající se chaperony, které usnadňují proces skládání.



*Obrázek 2-3- Skládání (https://commons.wikimedia.org/wiki/File:Protein_folding.png)*

**Exprese in vitro**

Exprese in vitro je prvním typem exprese proteinů. Při expresi in vitro jsou veškeré potřebné molekuly potřebné pro výrobu proteinu extrahovány z buněk (typicky Escherichia coli) a samotná exprese poté probíhá ve zkumavce. Takováto exprese probíhá velice rychle (řádově hodiny) a je vhodná pro výrobu proteinů náchylných na rychlou degradaci. Další výhodou je, že i proteiny, které by bylo pro živou buňku toxické lze takovýmto způsobem vyrobit. Tato metoda je také vhodná pro výrobu malého množství různých proteinů a hodí se tak pro výrobu proteinů pro vědecké účely. Nevýhodou této metody je poté to, že je nákladná a nelze s ní vyrábět proteiny v dostatečně velkém měřítku pro průmyslovou výrobu

**Exprese in vivo**

Při tomto typu exprese je do živých buněk vložen vektor, který obsahuje gen pro výrobu zvoleného proteinu. Následuje kultivace, při které probíhá rozmnožování těchto buněk. Poté se do prostředí přidá látka, která spustí transkripce a translaci proteinů. Buňky jsou následně lyzovány, což znamená že jejich buněčné membrány jsou rozloženy, což vede ke smrti buněk. Nakonec je provedena purifikace proteinu, což je proces, při kterém je z látky vzniklé lýzou buněk izolován pouze požadovaný protein. K tomuto typu exprese se používají buňky e. coli, kvasinky, buňky hmyzí, či buňky savčí. e. coli je ovšem pro svou jednoduchost, cenovou dostupnost a vysokou rychlost růstu nejpoužívanější. Výhodou

tohoto typu exprese je dobrá cenová dostupnost a jednoduchost výroby velkého množství proteinů. Nevýhodou této metody je to, že exprese probíhá v živých buňkách odlišných od originálního organismu, a tak je výroba toxických proteinů anebo eukaryotických proteinů při použití e. coli problematická. Řešením je právě použití hmyzích či savčích buněk. Takováto výroba je poté bohužel cenově náročnější a pomalejší.

## 2.2 Příčiny nerozpustnosti proteinů

Jak je zřejmé z předchozího textu, proces výroby proteinů je složitý a je mnoho věcí, které se při něm mohou nezdařit.

Jednou z příčin je tendence proteinu k agregaci. Jednotlivé řetězce se mohou spojovat do větších shluků, což jim zamezuje normálně fungovat. Tyto shluky mohou navíc být pro buňku toxické. Tendence proteinů agregovat zároveň souvisí s fází skládání proteinů. I složený a neagregující protein obsahuje místa, kterými by se mohl navázat na ostatní proteiny a agregovat. V takovémto proteinu jsou ovšem tyto místo schovány uvnitř jeho struktury a jsou tak pro ostatní proteiny nedostupné. Pokud tento protein budeme syntetizovat v jiném než jeho originálním prostředí, může docházet k pomalejšímu skládání, což vede k delšímu odhalení těchto míst a může zvýšit tendenci proteinu agregovat.

Při výrobě proteinu může dále vlivem nepříznivé teploty či pH docházet k nesprávnému skládání proteinu, což vede k výrobě nefunkčních proteinů. Mnoho proteinů také ke svému správnému složení vyžaduje chaperony anebo kofaktory. Chaperony jsou proteiny které pomáhají při skládání proteinů, často se na ně i naváží, ale ve složeném proteinu se již nevyskytují. Kofaktory jsou naopak neaminokyselinové struktury.

## 2.3 Zdroje dat

Pro tvorbu dobře fungujícího nástroje je nejprve důležité zajistit si dostatečné množství vhodných dat a z těchto dat poté získat datové sady obsahující veškeré relevantní informace pro trénování a testování tohoto nástroje. Pro tyto účely byly použity databáze eSol (1), SWISS-PROT (2), PDB (3) a Target Track (4).

**eSol**

Databáze eSol obsahuje cca 3200 proteinů bakterie e. coli. Každý protein byl syntetizován in vitro pomoci metody PURE (5). Při syntetizaci nebyli použity žádné chaperony. Rozpustnost proteinů byla poté testována pomoci centrifugace. Přibližně 1200 proteinů bylo poté označených jako nerozpustné. V roce 2012 provedli autoři dodatečné zvětšení databáze, kdy z 1200 nerozpustných proteinů vybrali 800 cytoplazmatických proteinů, kde zbylých 400 proteinů se v cytoplazmě nevyskytovalo (například šlo o proteiny vyskytují se v buněčné membráně). Těchto 800 proteinů se poté pokoušeli znovu vyrobit pomoci metody PURE a testovali, zda přidání tří hlavních typů chaperony ovlivní rozpustnost

těchto proteinů. Zjistili, že u dvou třetin proteinů došlo ke zvýšení rozpustnosti o více než 50 % a pouze u 3 % proteinů došlo ke zvýšení rozpustnosti menší než 20 %.

**SWISS-PROT**

SWISS-PROT je spravovaná databáze proteinů a jejich aminokyselinových sekvencí. Autoři se snaží o minimální redundanci, maximální integraci s ostatním databázemi a maximální anotaci všech obsažených proteinů (popis jejich funkce, struktury atd.). Databáze byla vytvořena roku 1987 v *Department of Medical Biochemistry of the University of Geneva* a je spravována ve spolupráci s *Department and the European Molecular Biology Laboratory* (EMBL)

**PDB**

On-line databáze *Protein Data Bank* se zabývá 3D strukturami proteinů, nukleových kyselin a jiných makromolekul. Databáze v současnosti obsahuje 129942 záznamů. Organizace *The Worldwide PDB* (wwPDB) spravuje tuto databázi a zajišťuje její bezplatnou dostupnost pro všechny uživatele kdekoliv na světě.

**Target Track**

Databáze Target Track je databází shromažďující experimentální výsledky snažící se získat strukturu proteinů vybraných organizací PSI (Protein Structure Initiative) a ostatních velkých biologických projektů zabývajících se strukturou proteinů. Obsahuje laboratoř, kde byl experiment proveden, zdrojový organismus proteinu a jeho doménu, protokoly použité při syntéze proteinu a výsledky jednotlivých trialů.

Tato databáze vznikla jako výstup projektu PSI, který byl rozdělen na několik fází:

- PSI-1 (2000–2005)
- PSI-2 (2005–2010)
- PSI: Biology (2010–2015)

Přestože projekt už skončil, databáze je stále aktualizována. V současné době obsahuje databáze cca 330 tisíc záznamů.

V Evropě existovaly podobné projekty s názvy SPINE (2003–2006) a SPINE2 (2006–2009). Databáze s výsledky těchto projektů bohužel neexistují.

## 2.4 Existující nástroje

Existuje mnoho nástrojů zabývajících se predikcí rozpustnosti proteinů. V této kapitole se budeme zabývat nástroji a metodami Wilkinson-Harrison(6), Idicula-Thomas (7), PROSO (8), SOLpro (9), SCM (10), PROSO II (11), ESPRESSO (12), ccSol (13), ccSol omics (14), CamSol (15) a Solubis (16). Nástroje můžeme rozdělit na tři generace.

### 2.4.1 Nástroje první generace

Do první generace patří nástroje, kde autoři prvně vybrali určité příznaky a poté za pomoc statistických metod prováděli predikci rozpustnost.

**Wilkinson-Harrison**

Na základě sekvencí 81 proteinů se známou informací o rozpustnosti vytvořili autoři pěti-parametrový model (rovnici) pro predikci rozpustnosti. Tento model bral v úvahu:

- průměrný náboj
- obsah aminokyselin tvořících otočky – proteiny s velkým počtem otoček se hůře skládají
- obsah cysteinů - E. coli neumí vytvářet disulfidické můstky, proto se řada savčích proteinů s velkým podílem cysteinů nesloží správně.
- obsah prolinů – proline má výrazný vliv na rychlost skládání proteinu.
- hydrofilitu – ovlivňuje stabilitu proteinu
- celkový počet reziduí

Autoři zjistili na základě jednoduché statistiky, že právě těchto pět parametrů nejlépe rozlišuje rozpustnost.
Datová sada použitá při tvorbě této metody obsahuje 54 solubilních a 27 nesolubilních proteinů z různých zdrojů (člověk, kuře, kvasinka, bakterie apod.).

### 2.4.2 Nástroje druhé generace

Nástroje druhé generace se liší od první generace tím, že místo statistické metody využívají pro zpracování příznaků různé metody strojového učení anebo jejich kombinace. Zvlášť populární se zdá být metoda SVM (Simple Vector Machine).

**Idicula-Thomas**

Tvůrci nástroje z aminokyselinové struktury proteinu a z nich vypočítaných vlastností vytvořily soubor rysů. Například použily délku proteinů, GRAVY, alifatický index, celkový náboj atd. Různé kombinace těchto rysů poté testovali jako vstupní vektory pro SVM. Autoři ale používali pouze velice malou datovou sadu 62 rozpustných a 130 nerozpustných proteinů, kterou dále rozdělili na trénovací

sadu (41 rozpustných a 87 nerozpustných proteinů) a testovací sadu (21 rozpustných a 43 nerozpustných proteinů). Přesnost tohoto nástroje se pak pohybovala mezi 66 a 72 %, v závislosti na volbě rysů.

**PROSO**

Jako vstupy pro SVM používají autoři frekvenci výskytu jednotlivých aminokyselin, frekvenci výskytu dvojic aminokyselin a frekvenci výskytu trojic aminokyselin v sekvenci zkoumaného proteinu. Protože by ale u dvojic a trojic došlo k velkému množství možných kombinací, shlukovali autoři nástroje aminokyseliny s podobnými vlastnostmi do shluků a použily poté frekvence těchto shluků. Výsledky těchto tří SVM byly poté zpracovány pomoci naivní bayesovské klasifikace. Datovou sadu vytvořili autoři pomoci databáze TargetDB (17) (předchůdce Target Tracku) tak, že jako rozpustné označili všechny proteiny které dosáhli stavu *soluble* a jako nerozpustné ty, které dosáhli alespoň stavu *expressed,* ale nikoliv *soluble*. Tyto proteiny poté přezkoumali pomoci nástroje TMHMM (18) a vyloučili proteiny, při kterých byla predikována přítomnost transmembránových segmentů. Pomoci nástroje CD-HIT (19) poté zbylé proteiny poshlukovali s použitím 50 % hranice podobnosti. Výslednou datovou sadu autoři poté rozdělili na proteiny obsahující jednu doménu a na proteiny obsahující více domén. Dělení provedly tak, že všechny proteiny delší než 150 aminokyselin označili za vícedoménové. Získaly tak dvě datové sady:
Jednodoménovou (2 x 3117 proteinů)
Vícedoménovou (2 x 3983 proteinů)
Autoři nástroje poté ohlásili přesnost 71,7 % avšak autoři prediktoru SOLpro při testování nad jejich datovou sadou ukázali přesnost prediktoru PROSO jen 59,285 %.

**SOLpro**

Podobně jako PROSO se SOLpro skládá ze dvou fází. V první fázi je vykonáno 20 SVM klasifikátorů, každý nad jinou sadou rysů. Jako rysy jsou použity jednak frekvence výskytu jednotlivých aminokyselin, dvojic a trojic, podobně jako je tomu u nástroje PROSO, délka proteinu, GRAVY, aliphatický index, náboj proteinu, molekulární váha, náchylnost proteinu na tvorbu, poměr alpha residuí, beta residuí a počet domén. Ve druhé fázi jsou výsledky první fáze společně s délkou proteinu zpracovány opět pomoci SVM. Autoři vytvořili vlastní datovou sadu SOLP, která byla vytvořena převážné z databází PDB a TargetDB a obsahovala 17408 proteinů. Pro testování byla využita 10násobná křížová validace, kde každá část obsahovala vyvážené množství rozpustných a nerozpustných proteinů. Autoři nástroje udávají přesnost 74,15 %.

**SCM**

Tento nástroj využívá novou metodu predikce, kterou jeho tvůrci nazývají SCM (Scoring Card Method). Autoři prvně vytvořili vlastní sadu Sd957, kterou získaly přímo z literatury. Tuto sadu poté

náhodně rozdělili na 766 trénovacích a 191 testovacích prvků. Následně autoři použili trénovací sadu na tvorbu bodové matice. Prvně rozdělili sadu na rozpustné a nerozpustné proteiny. Poté si pro obě části sady spočítaly poměr všech 400 di-peptidů vůči ostatním di-peptidům. Poté odečetly poměry u nerozpustných proteinů od poměrů u rozpustných proteinů. Nakonec všechny hodnoty normalizovali do rozsahu 0–1000. Podle této matice se poté pro každý protein spočítá skóre a pokud přesahuje danou hranici, považuje se protein za rozpustný, jinak za nerozpustný. Autoři uvádějí přesnost kolem 83 %. Nástroj dodatečně testovali na sadách SOLP, kde byla přesnost 59,99 % a PROSO, kde byla přesnost 64,5 %.

**PROSO II**

Tento nástroj je vylepšenou verzí nástroje PROSO a funguje tedy na velice podobném principu. Pouze místo SVM využívá Parzenovo okno a logistickou regresi. Autoři také použili výrazně větší datovou sadu obsahující 82299 proteinů. Tuto sadu poté ještě poshlukovali na 90 % a vybalancovali. Nástroj nakonec dosahoval přesnosti 71 %.

**ESPRESSO**

ESPRESSO je v podstatě dvěma rozdílnými nástroji v jednom, jelikož pracuje se dvěma metodami predikce, mezi kterými si může uživatel vybrat. První je predikce na základě vlastností proteinu, kde autoři nejprve definovali 437 rysů na základě jak samotné sekvence, tak i její struktury a poté použili opět SVM na zpracování těchto rysů. Druhou metodou je predikce podle vzorce, kde autoři hledali v proteinu sekvenční vzory, které se vyskytovali buď pouze u rozpustných, nebo nerozpustných proteinech. Autoři využili datovou sadu, kterou používaly v jejich předchozích pracích. Tuto sadu rozdělili na dvě části. Jednu, kde byla rozpustnost zjištěna pouze jedním experimentem (dataset_S) a druhou část, kde byla rozpustnost otestována dvěma, nebo více experimenty (dataset_M). Sady byly poté shlukovány nástrojem CD-HIT, ale nebyly vyrovnané (obsahovaly přibližně dvojnásobné množství nerozpustných proteinů). Na této datové sadě dosáhl nástroj pomoci první metody přesnosti 68 % a pomoci druhé metody 63 %.

**ccSol**

Při tvorbě nástroje tvůrci prvně pro každou sekvencí vypočítaly 28 různých fyzikálně-chemických vlastností a tyto poté zredukovali testováním na 6 rysů pro SVM. Těmi jsou vinutí proteinu, hydrofobicita, hydrofilita, neuspořádanost proteinu, beta otočky a alpha spirály. Autoři používají datovou sadu vycházející z databáze eSol, která obsahuje 3043 proteinů. Autoři bohužel neuvádějí přesnost nástroje

### 2.4.3 Nástroje třetí generace

Výsledkem nástrojů třetí generace není pouze predikce, zda bude protein rozpustný, ale tvorba profilu rozpustnosti proteinu. Tento přístup umožňuje predikovat přímo vlivy mutací na rozpustnost proteinu. Nevýhodou je, že nástroje často vyžadují znalost nejen aminokyselinového řetězce protenu, ale i jeho struktury.

**CamSol**

Motivací je zvýšit rozpustnost léků ve formě *antibodies proteinů*, které mají tendence agregovat ve větších koncentracích, a proto se zaměřuje primárně na nalezení a rozrušení agregujících regionů. Jako vstup nástroj používá strukturu proteinu. Prvně vypočte profil agregace podobným principem, jako nástroj Zyggregator (z tendence pro alpha šroubovice, beta skládaný list, hydrofobicita a náboje). Poté se provedou strukturní korekce na základě výskytu konkrétní části profilu v jádře nebo na povrchu proteinu. Na základě výsledků jsou poté vybrána vhodná místa pro mutace. Autoři použili datovou sadu 56 mutací na 19 proteinech. Nástroj byl porovnán s výsledky nástrojů eSol a PROSSO II a CamSol dopadl nejlépe (54/56 bylo predikováno správně).

**Solubis**

Tento nástroj využívá sekvence a struktury proteinu. Nástroj nejprve vyhledá *Aggregation Prone Regions* pomoc nástroje TANGO (20). Dále se analyzuje vliv jednotlivých APR regionů na stabilitu proteinu podle jejich umístění v jádře, nebo mimo ně. Nakonec jsou jednotlivé APR regiony skenovány na možné mutace a jsou posuzovány z pohled změny náchylnosti k agregaci a změny ve stabilitě. Výstupem nástroje je poté graf ukazující nejperspektivnější mutace a jejich souvislost se změnou agregace a změnou stability.

**ccSol omics**

Nástroj ccSol omics je novější verzí ccSol. Používá algoritmus nástroje ccSol, ale rozpustnost nepočítá jako jednu hodnotu pro protein, ale pracuje s posuvným okénkem a vytváří tak spíše profil proteinu. Ke zpracování profilu na jednoznačný výsledek používá nástroj neurální sítě. Datová sada vychází z databáze Target Track a obsahuje cca 37 tisíc proteinů. Přesnost nástroje uvádějí autoři 79 %.

# 3 Praktická část

Jak můžeme vidět v kapitole 2.4, existuje mnoho nástrojů pro predikci rozpustnosti proteinů. Téměř všechny nástroje používají pro predikci SVM nebo jiné formy strojového učení a zdá se být trendem rozdělit zpracovávání výsledků do dvou fází a mezi několik metod. Můžeme také vidět, že většina nástrojů se zabývá buď vlastnostmi proteinu anebo hledáním určitých vzorů v jeho řetězci. Dále je zřejmé, že čím jsou nástroje starší, tím menší datové sady jejich autoři používají. Navíc jsou často datové sady získávány z podobných zdrojů, a tak zde vyvstává i otázka, zda jsou takovéto datové sady reprezentativní pro obecnou predikci rozpustnosti proteinů. Našim cílem je tedy vytvoření obsáhlejší datové sady, než používá jakýkoliv z popsaných nástrojů a otestování co nejvíce potenciálních rysů pro tvorbu prediktoru.

## 3.1 Tvorba datových sad

Pro potřebu testování různých příznaků pro strojové učení a testování samotného nástroje byly vytvořeno několik datových sad.

**Datová sada eSol**

Na základě databáze eSol byly vytvořeny dvě testovací sady. Databáze eSol obsahuje informace o rozpustnosti proteinů, ovšem neobsahuje informace o sekvencích proteinů. Místo sekvencí obsahuje pouze ECK identifikátory, a tak bylo třeba dohledat proteinové sekvence pomoci databáze SWISS-PROT. Takto připravená sada poté obsahuje 3133 proteinů, kde jsou informace o míře rozpustnosti proteinu místo pouze binární informace, zda je protein rozpustný, nebo ne. Databáze eSol také obsahuje u 773 proteinů experimentální informace o změně rozpustnosti proteinů při přidání chaperonů, a tak byla pro účely testování příznaků vytvořená druhá datová sada, obsahující pouze tyto proteiny.

**FitSet**

Na základě databáze Target Track vzniklo několik datových sad souhrnně označovaných jako FitSet. Jelikož databáze neobsahuje přímo binární informace o rozpustnosti proteinů, ale pouze trialy a jejich výsledky, byla rozpustnost proteinů vypočtena jako poměr rozpustných triálů vůči nerozpustným. Binární data o rozpustnosti byla poté získána tak, že pokud byl poměr větší než 50 %, tak byl protein označen za rozpustný, jinak za nerozpustný. Databáze byla prvně rozdělena na solubilní a nesolubilní část. Takto připravená databáze byla poté shlukována pomoci nástroje CD-HIT postupně na podobnosti 90 %, 80 %, 70 %, 60 % a 50 %. Nakonec byly datové sady vyrovnány náhodným výběrem postupně z rozpustných proteinů a poté z nerozpustných tak dlouho, dokud se jakákoliv část nevyčerpala. Vznikly tak tyto datové sady

- FitSet0 – neshlukovaná a nevyrovnaná sada vycházející z databáze (272662 rozpustných a 92314 rozpustných proteinů)
- FitSet9 – vyrovnaná a shlukovaná sada s podobností 90 % (2x 61247 proteinů)
- FitSet8 – vyrovnaná a shlukovaná sada s podobností 80 % (2x 58826 proteinů)
- FitSet7 – vyrovnaná a shlukovaná sada s podobností 70 % (2x 56204 proteinů)
- FitSet6 – vyrovnaná a shlukovaná sada s podobností 60 % (2x 52625 proteinů)
- FitSet5 – vyrovnaná a shlukovaná sada s podobností 50 % (2x 47202 proteinů)

## 3.2 Volba příznaků

Na základě teoretických znalostí a výsledků předchozích nástrojů byly vybrány příznaky z několika různých kategorií pro testování. Veškeré testování probíhalo za pomocí skriptů v jazyce Python pro spouštění jednotlivých nástrojů a práci s nimi a za pomoci scriptů v jazyce R pro tvorbu grafů a statistické zpracování dat.

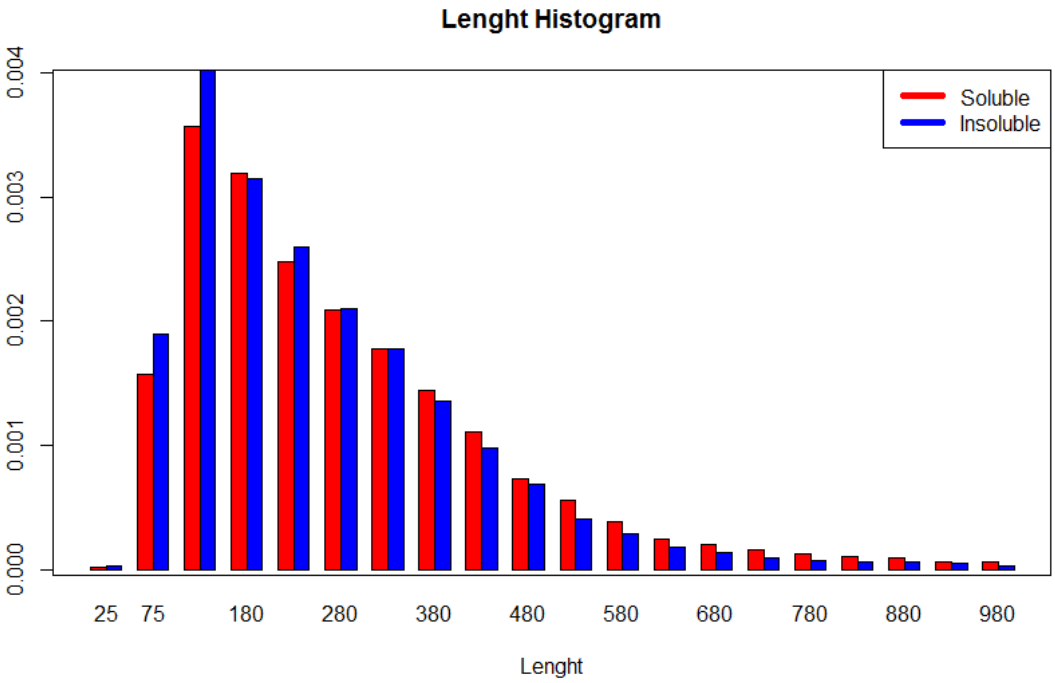
### 3.2.1 Vlastnosti proteinů

Dle dosavadních výsledků nástrojů se zdá, že vlastnosti proteinů mají vliv na jejich rozpustnost. Pro všechny proteiny v sade FitSet5 byly tedy vypočítány tyto vlastnost:

- Délka sekvence proteinu
- GRAVY
- Náboj
- Aliphatický index
- Izoelektrický bod

**Délka**

První testovanou vlastností byla délka proteinu. Ta je jednoduše vypočítána ze sekvence proteinu tak, že protein je tak dlouhý, kolik obsahuje aminokyselin.



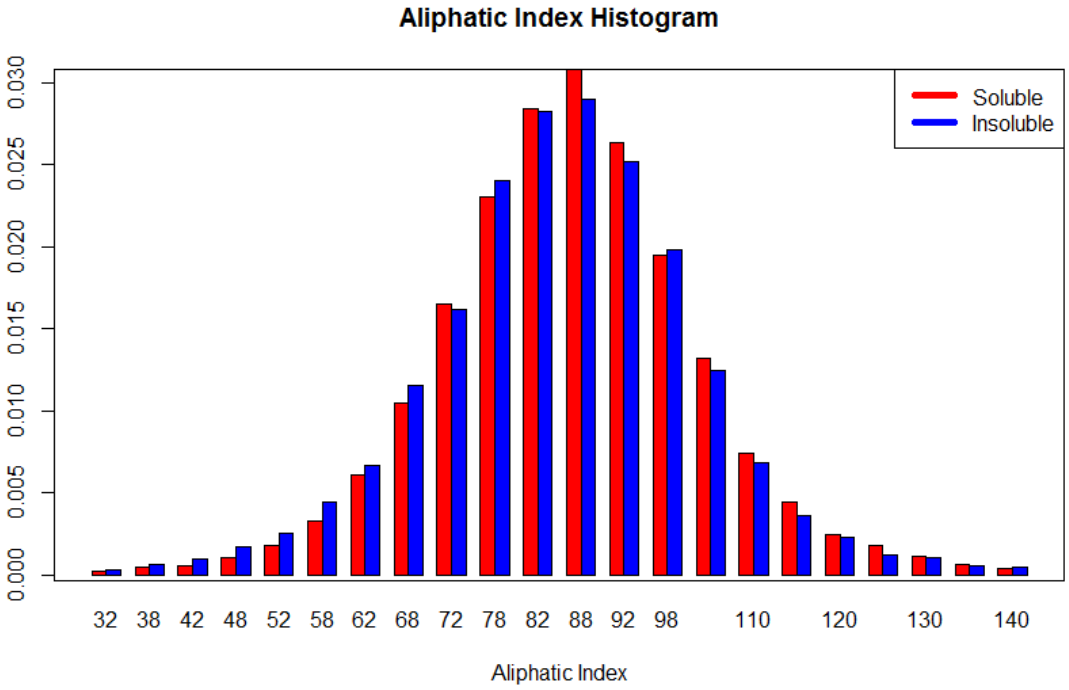

*Obrázek 3-1: Histogram poměru solubilních a nesolubilních proteinů a jejich délek*

### Aliphatický Index

Tato vlastnost byla vypočtena podle vzorce:

$$\text{Aliphatický Index} = \text{Ala} + 2.9 * \text{Val} + 3{,}9 * (\text{Ile} + \text{Leu})$$

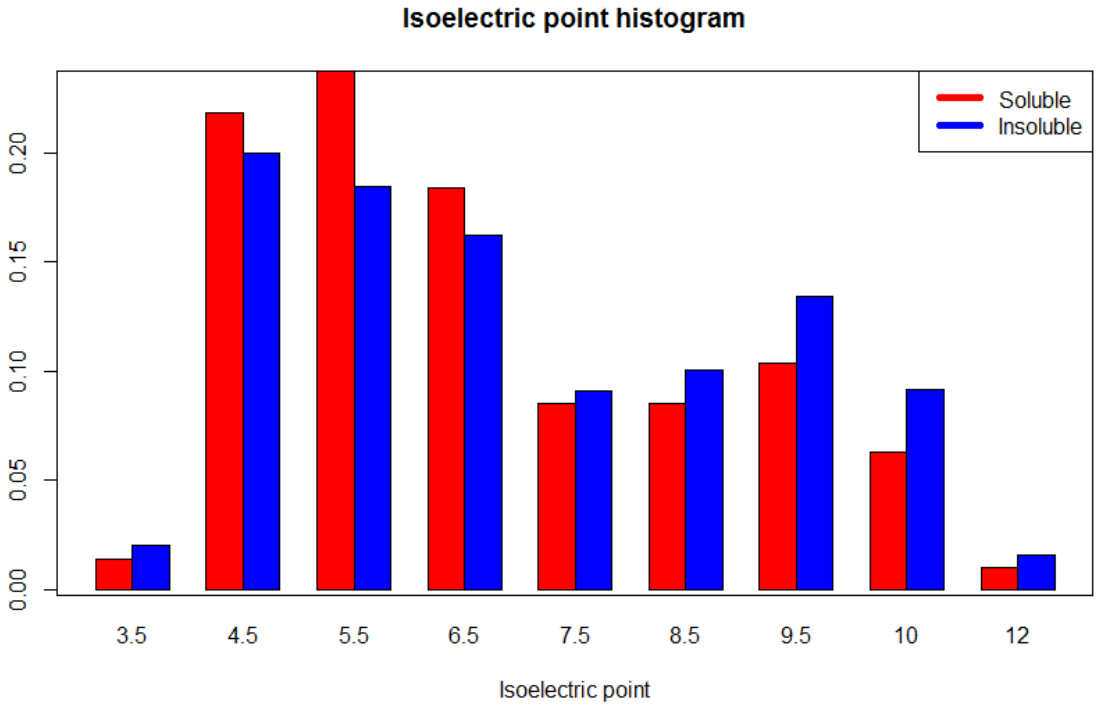
Kde *Ala*, *Val*, *Ile* a *Leu* jsou poměry aminokyselin Alaninu, Valinu, Izoleucinu a Leucinu v proteinu v procentech.




*Obrázek 3-2: Histogram poměru solubilních a nesolubilních proteinů a jejich Aliphatických indexů*

### Isoelektrický bod

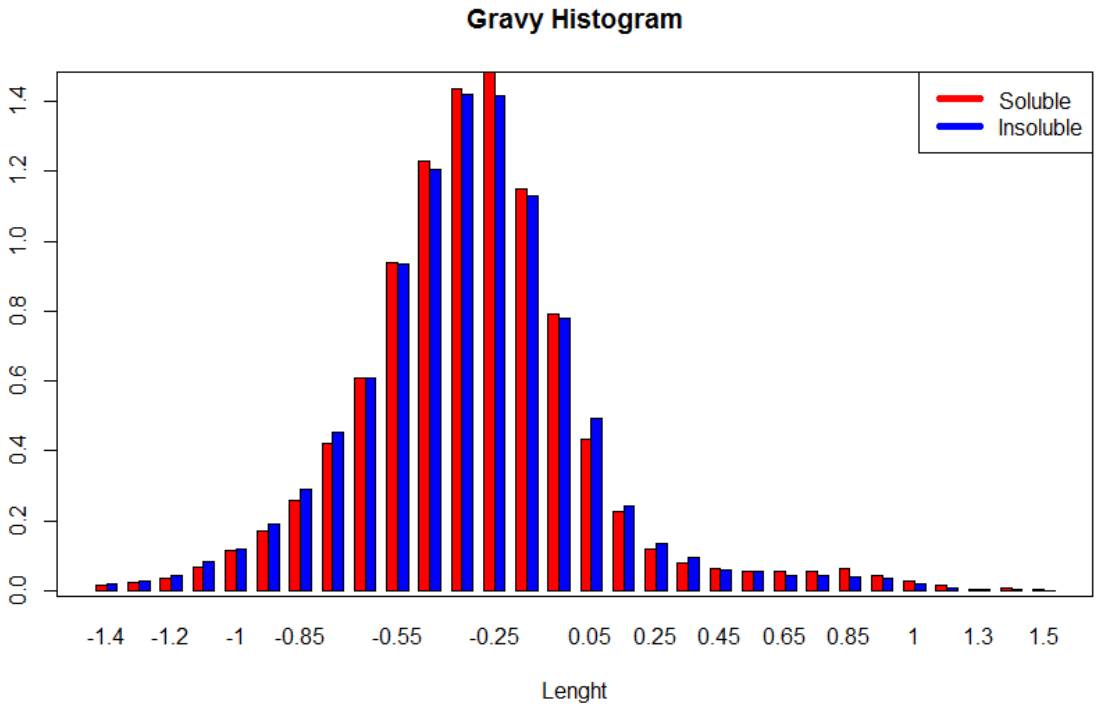
Isoelektrický bod je bod na stupnici pH, kde má protein nulový náboj. Byl určen numericky tak, že se po pH stupnici postupoval s krokem 0,1 a pro každé pH byl vypočítán náboj. Jako výsledek byl určen bod s nejnižším vypočítaným nábojem.




*Obrázek 3-3: Histogram poměru solubilních a nesolubilních proteinů a jejich Izoelektrických bodů*

### GRAVY

Vlastnost GRAVY byla vypočtena tak, že každé aminokyselině je přiřazena určitá numerická hodnota, všechny tyto hodnoty proteinu jsou poté sečteny a vyděleny délko proteinu.



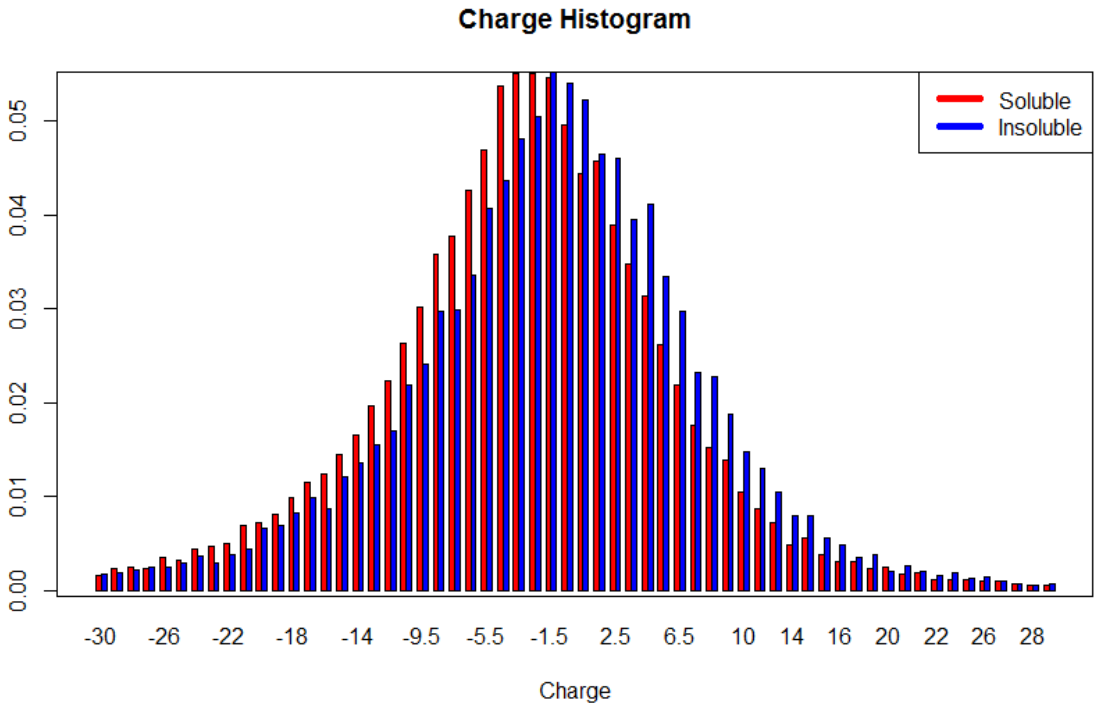

*Obrázek 3-4: Histogram poměru solubilních a nesolubilních proteinů a hodnoty GRAVY*

**Náboj**

K výpočtu náboje potřebujeme několik hodnot. Potřebujeme znát počty jednotlivých aminokyselin v proteinu, což lze lehce spočítat ze sekvence a potřebujeme znát pH, které jsme určili na konstantní hodnotu 7 (neutrální). Z těchto dat se dle následujícího vzorce provede výpočet náboje:

$$\text{Náboj} = -\frac{1}{1+10^{3.65-\text{pH}}} - \frac{\text{AspNumber}}{1+10^{3.9-\text{pH}}} - \frac{\text{GluNumber}}{1+10^{4.07-\text{pH}}} - \frac{\text{CysNumber}}{1+10^{8.18-\text{pH}}} - \frac{\text{TyrNumber}}{1+10^{10.46-\text{pH}}} + \frac{\text{HisNumber}}{1+10^{\text{pH}-6.04}} + \frac{1}{1+10^{\text{pH}-8.2}} + \frac{\text{LysNumber}}{1+10^{\text{pH}-10.54}} + \frac{\text{ArgNumber}}{1+10^{\text{pH}-12.48}}$$

Kde *AspNumber*, *GluNumber*, *CysNumber*, *TyrNumber*, *HisNumber*, *LysNumber* a *ArgNumber* jsou počty aminokyselin Kyselina asparagová, Kyselina glutamová, Cystein, Tyrosin, Histidin, Lysin a Arginin respektive.




*Obrázek 3-5: Histogram poměru solubilních a nesolubilních proteinů a jejich náboje*

**Závěr**

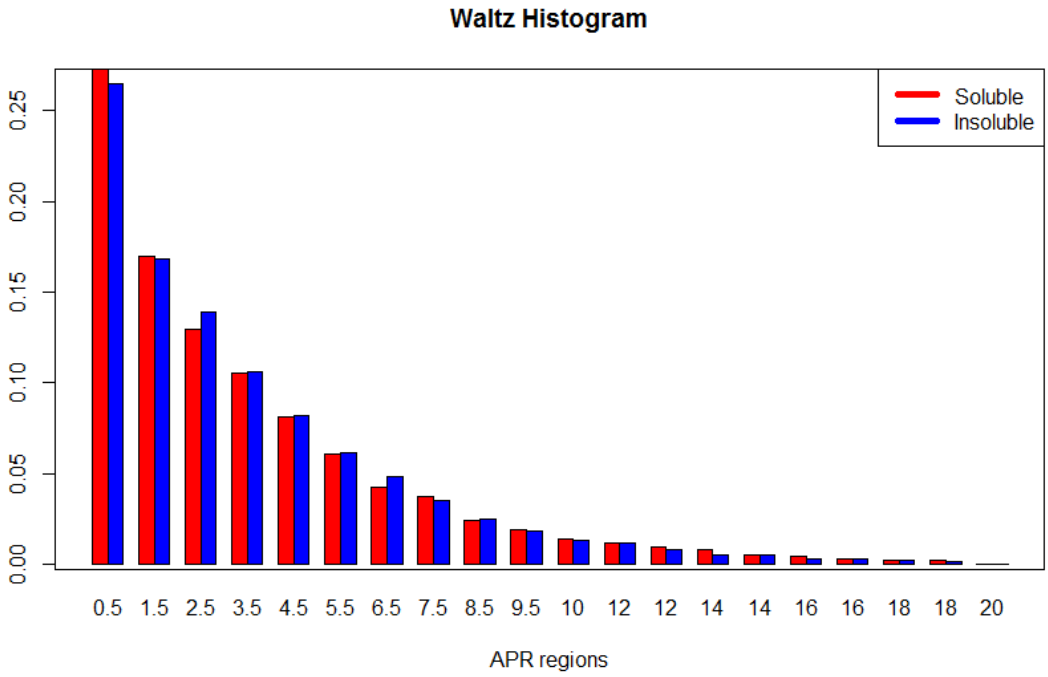
Bohužel se nepodařilo nalézt silné korelace mezi vlastnostmi a rozpustností proteinů, ale existují slabé korelace. Toto je nejlépe vidět na náboji a izoelektrickém bodě.

### 3.2.2 Agregace

Agregace proteinů by logicky měla ovlivňovat rozpustnost proteinů z toho důvodu, že agregovaný protein nefunguje správně a měl by se tedy dát označit za nesolubilní. Pro analýzu agregace proteinů jsme se rozhodli použít příznaky z nástroje, který jsme vyvinuli v předchozí práci. (21). Dále jsme se rozhodli otestovat vliv APR (*Aggregation Prone Region*) na agregaci. Tyto regiony byly hledány nástroji Tango (20), Zyggregator (22) , Waltz (23) a FitSeq (21).
Nad sadou FitSet5 byly testovány všechny nástroje s výjimkou Tanga a nad sadou eSol byly testovány nástroje Tango a Zyggregator.

**Waltz**

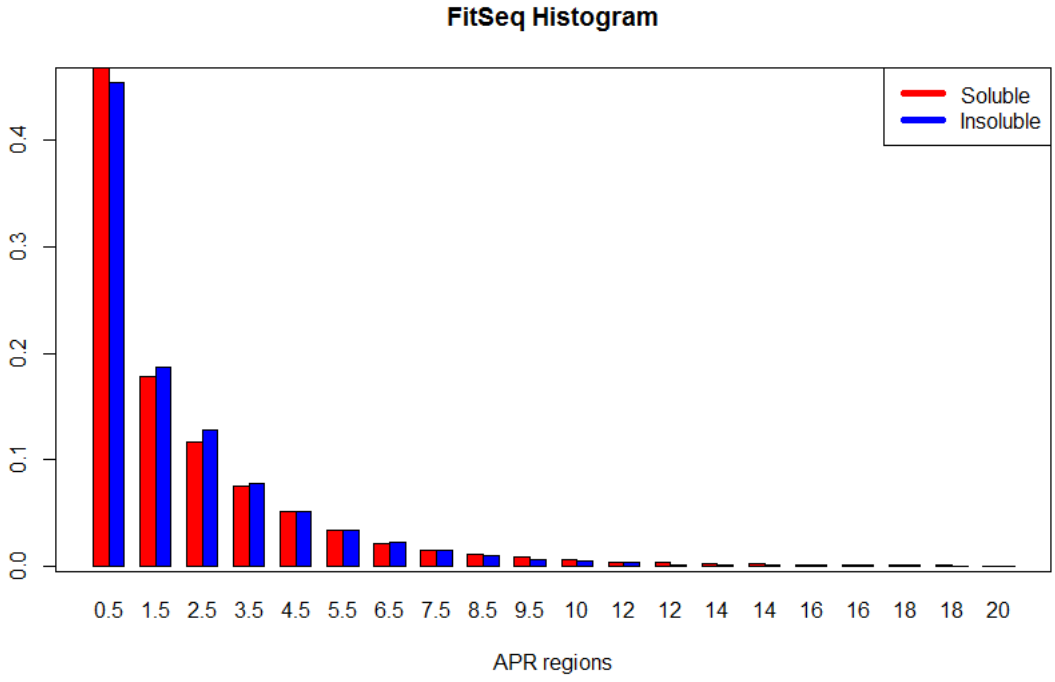
Nástroj pouze hledá APR. Byly testovány počet, průměrná síla a součet sil těchto regionů. Nepodařilo se najít korelace mezi APR a rozpustností (Korelační koeficient 0.01936721, viz obrázek 3-6).




*Obrázek 3-6: Histogram poměru solubilních a nesolubilních proteinů a počtu APR hledaných nástrojem Waltz na sadě FitSet5*

**FitSeq**

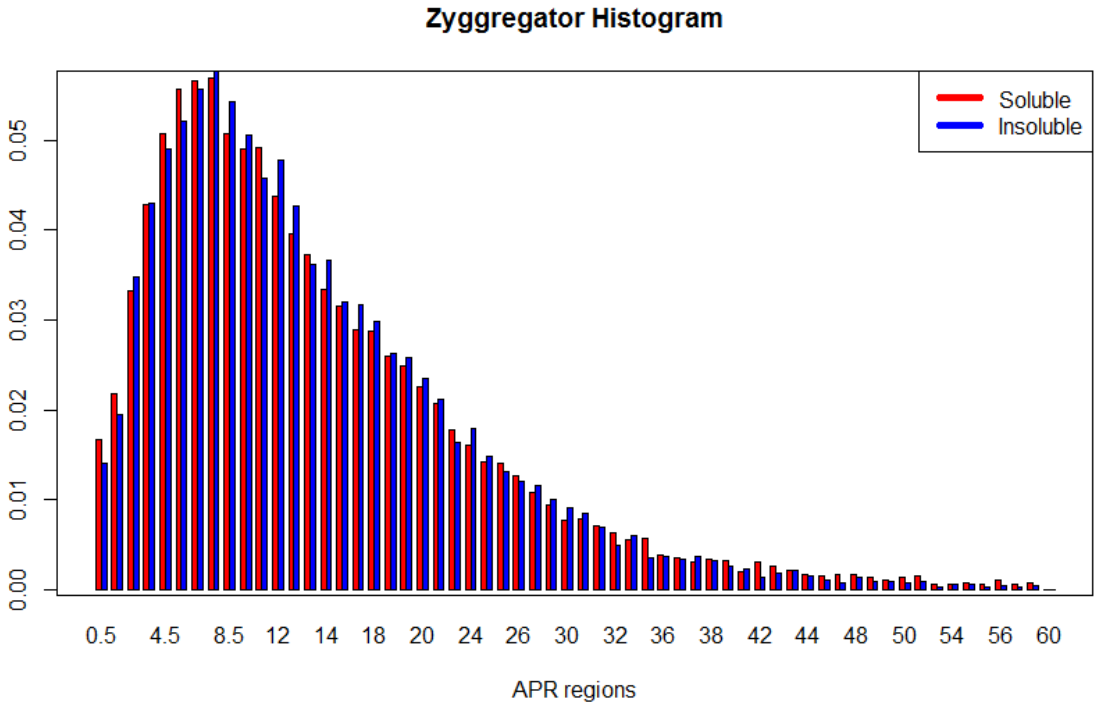
Nástroj hledá specifické sekvence aminokyselin. Testováno bylo, zda byla nalezena alespoň jedna sekvence a počet nalezených sekvencí. Každá taková sekvence poté byla považována za APR. Korelaci mezi počtem APR a rozpustností proteinu se opět nepodařilo prokázat (Korelační koeficient 0.02264467, viz obrázek 3-7).




*Obrázek 3-7: Histogram poměru solubilních a nesolubilních proteinů a počtu APR hledaných nástrojem FitSeq na sadě FitSet5*

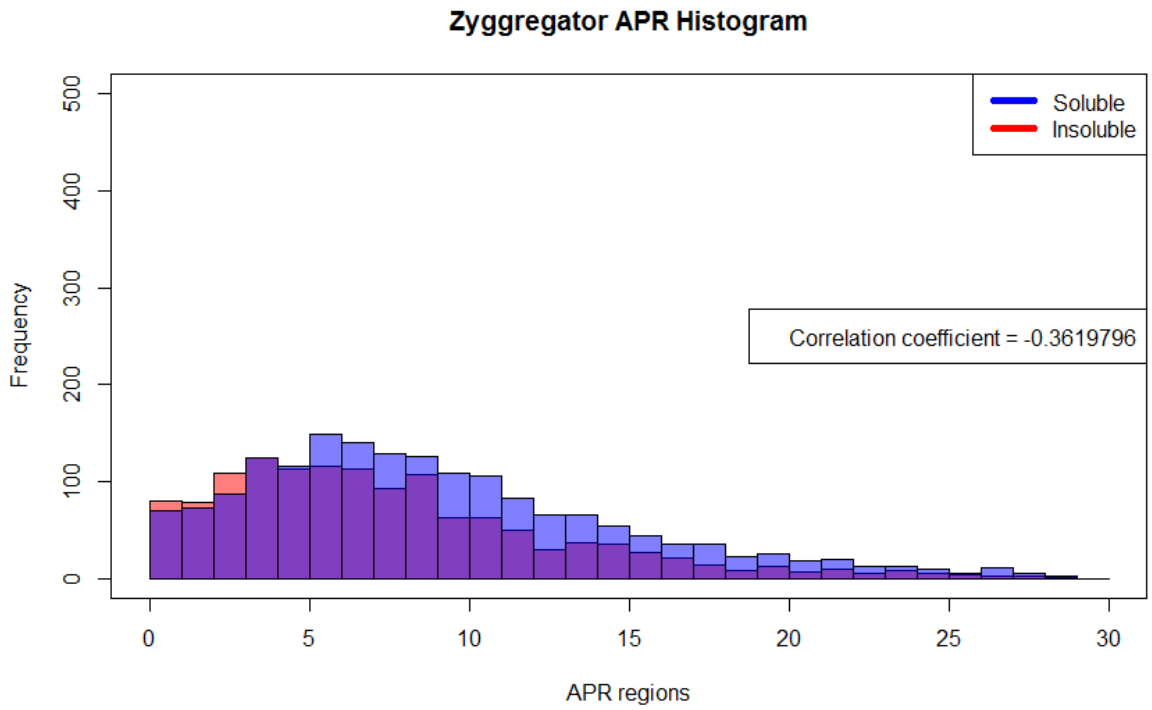
**Zyggregator**

Tento nástroj vytváří profil proteinu a vypočítává navíc vlastnosti jako hydrofobicita, pravděpodobnost výskytu beta skládaného listu, pravděpodobnost výskytu alpha šroubovice. Ani u jedné z těchto vlastností se bohužel nepodařilo na sadě FitSet5 prokázat korelaci s rozpustností proteinu. (korelační koeficient 0.02488598, viz obrázek 3-8).

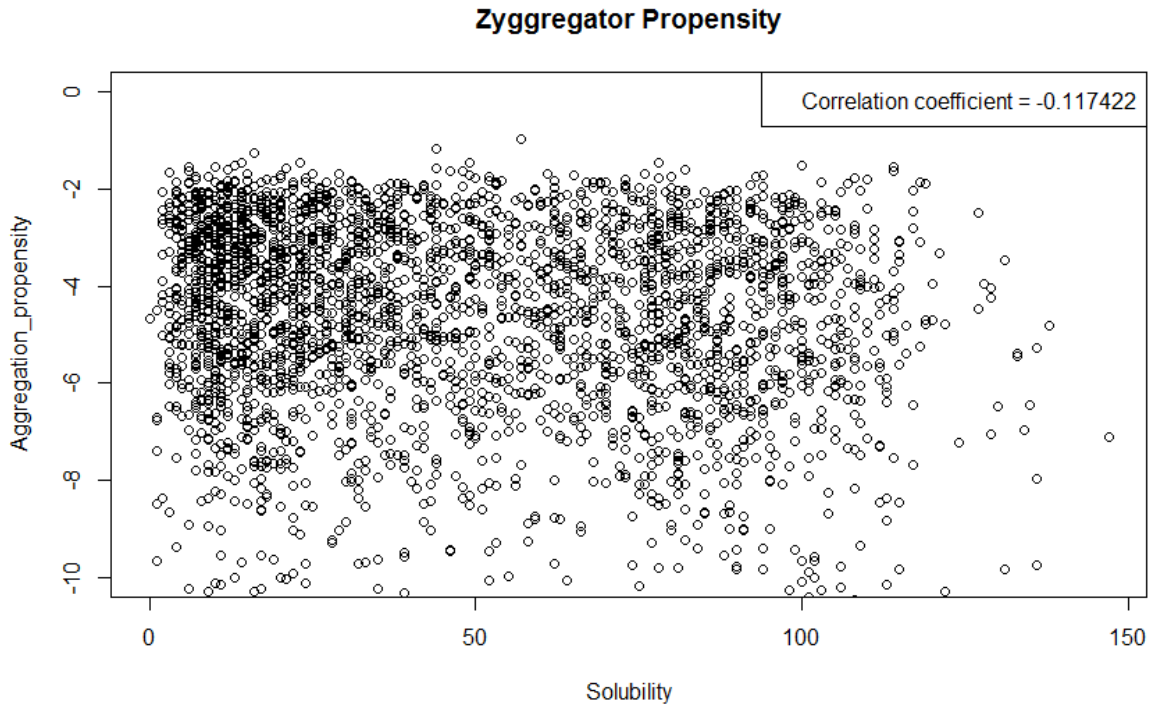



*Obrázek 3-8: Histogram poměru solubilních a nesolubilních proteinů a počtu APR hledaných nástrojem Zyggregator na sadě FitSet5*

Nástroj byl dále testován na sadě eSol. V tomto experimentu bylo testováno jak množství APR, tak samotná předpovídaná míra agregace proteinů (viz obrázky 3-9 a 3-10). Byla zjištěna slabá korelace mezi APR a rozpustností proteinů.



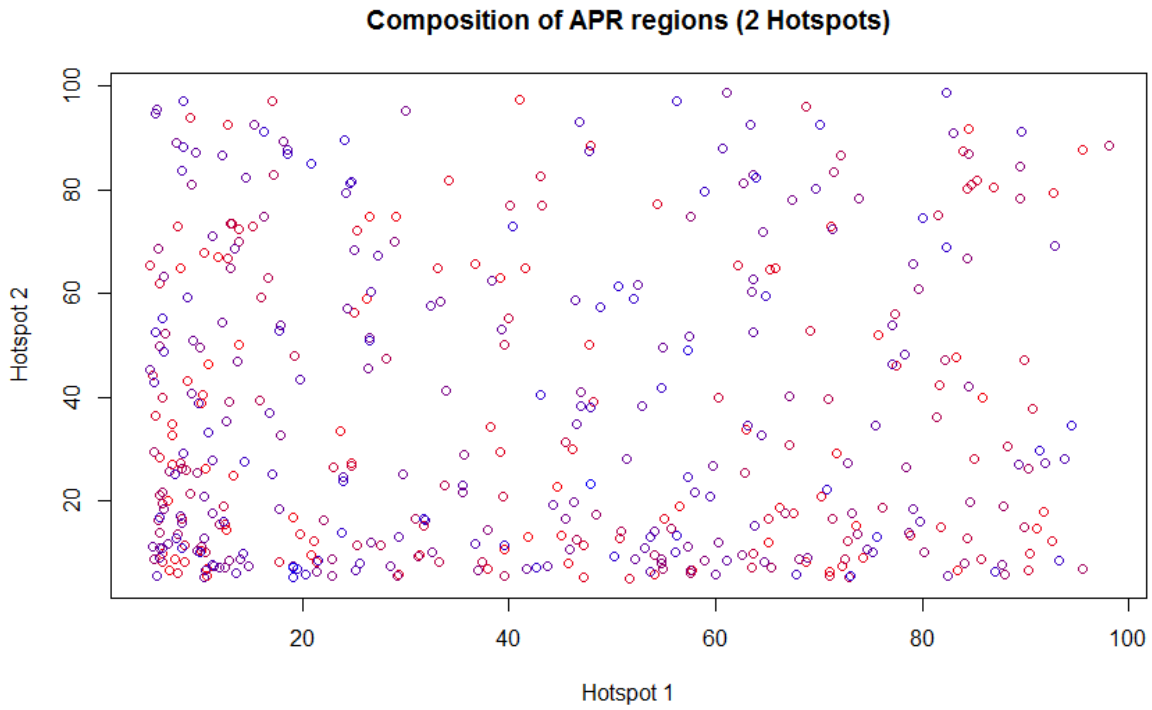

*Obrázek 3-9: Histogram poměru solubilních a nesolubilních proteinů a počtu APR hledaných nástrojem Zyggregator na sadě eSol*




*Obrázek 3-10: Vztah předpovídané tendence agregovat a rozpustnosti proteinu v sadě eSol*

**Tango**

Pomocí nástroje Tango byl proveden experiment odlišný od ostatních nástrojů. Pro všechny hotspoty byla na sadě eSol spočítána jejich síla a zkoumalo se, zda nějak koreluje síla hotspotů s rozpustností proteinů. U proteinů, které obsahují více než dva hotspoty jsou data upravena pro vizualizaci v grafu pomoci nástroje XY. Modrá barva, podobně jako u předchozích grafů, reprezentuje nerozpustné proteiny a červená rozpustné. Bohužel se nepodařilo najít žádné korelace. (viz obrázek 3-11 a DVD).



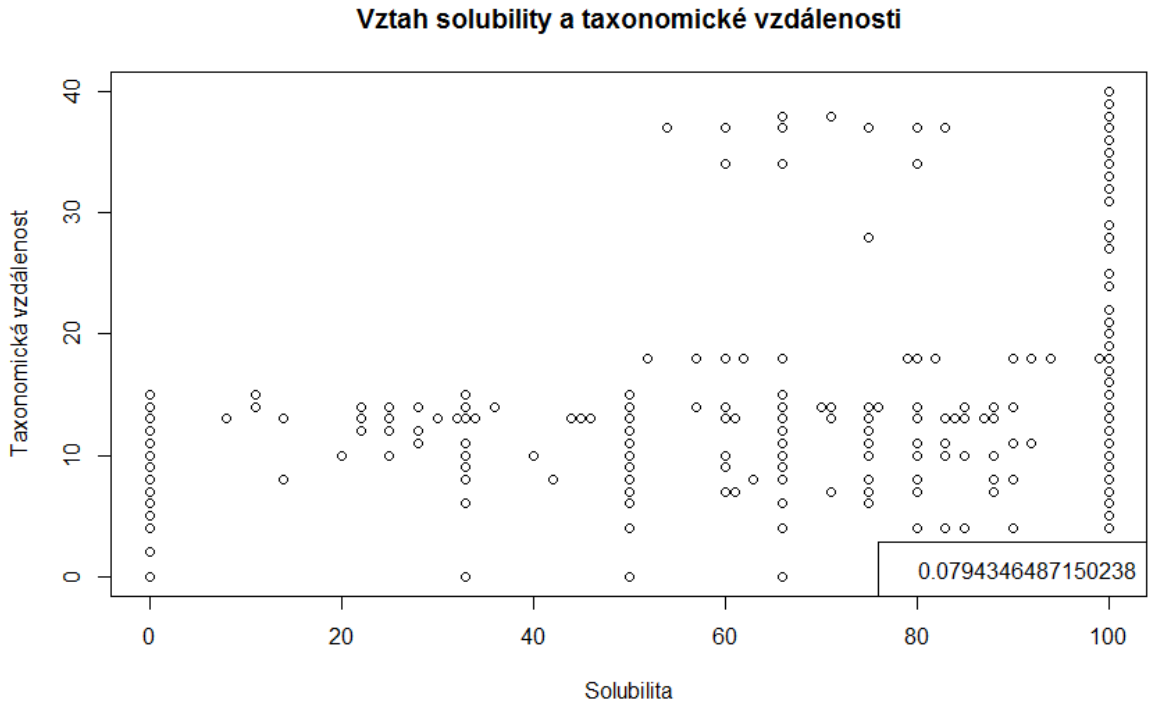

*Obrázek 3-11: Poměr síly APR hledaných nástrojem tango a rozpustnosti protenu pro proteiny se dvěma APR (modré rozpustné, červené nerozpustné)*

**Závěr**

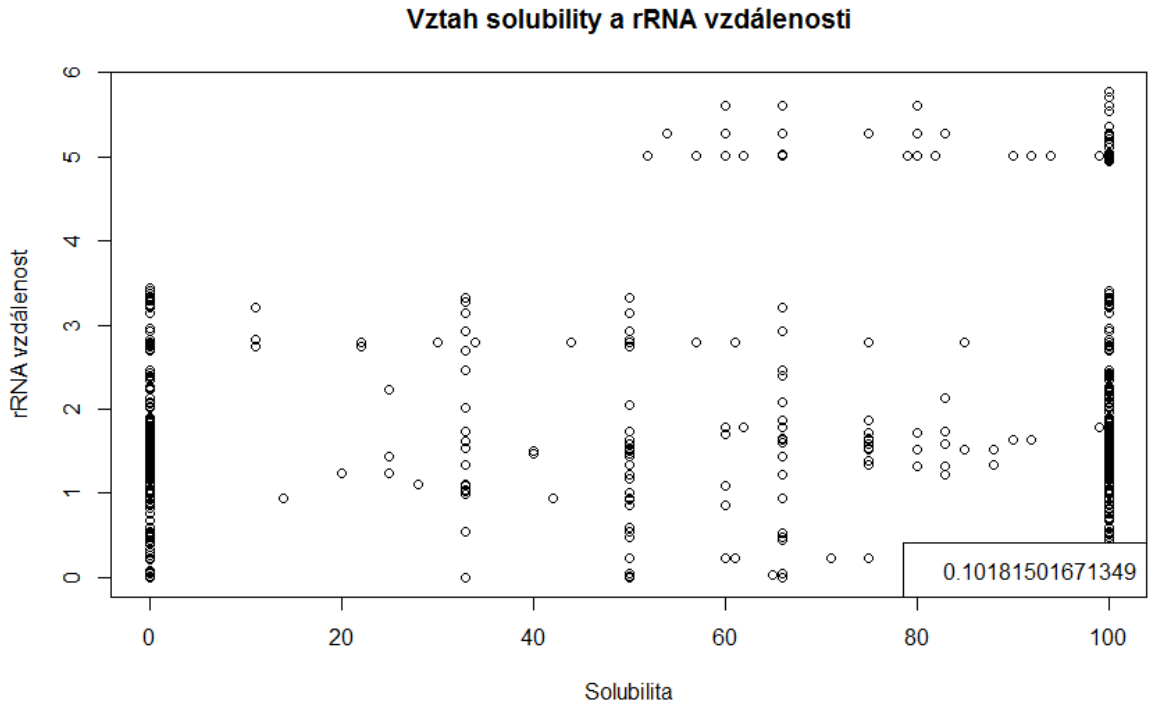
S výjimkou nástroje Zyggregator na datové sadě eSol se kupodivu nepodařilo prokázat ani slabou korelaci počtu a síly APR regionů a rozpustnosti proteinů.

### 3.2.3 Zdrojový organismus

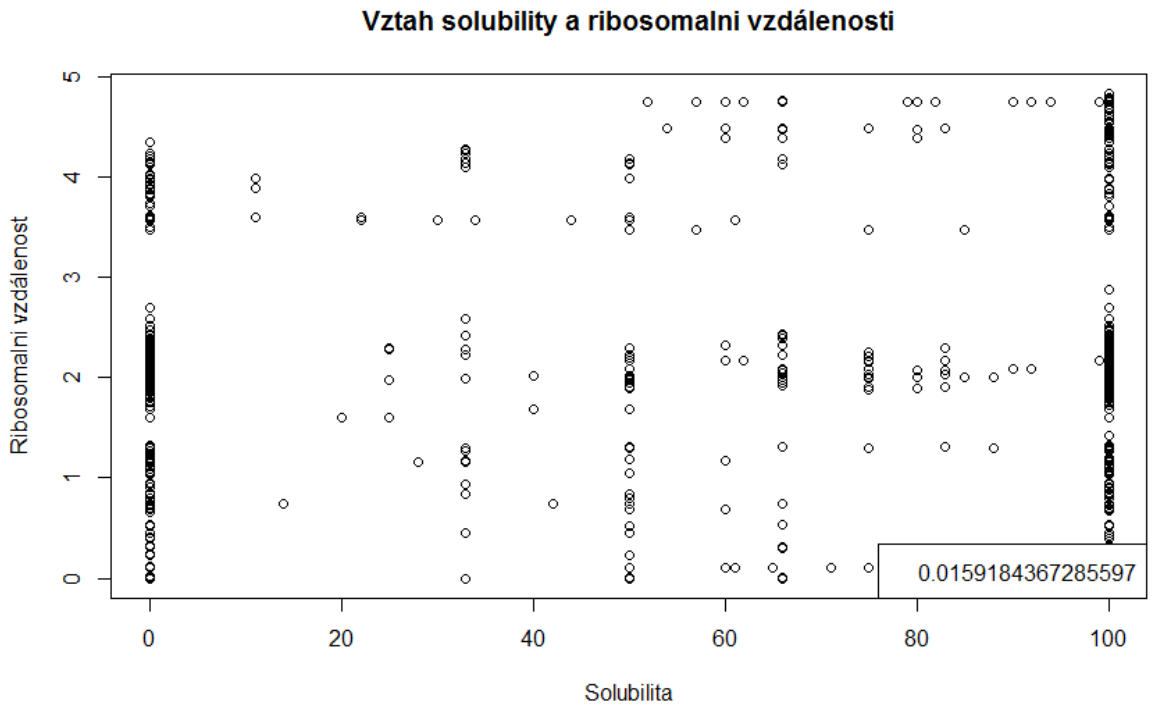
Rozhodli jsme se také otestovat vzdálenosti organismů vůči e. coli, která je nejčastěji používaná jako hostitelský organismus. Byly použity tři vzdálenosti: Taxonomická, fylogenetická podle rRNA a fylogenetická podle ribozomů. Tyto vzdálenosti byly poté testovány na eSol jako potenciální příznaky pro prediktor. Bohužel se ani u jedné z těchto vlastností neprokázala silná korelace s rozpustností proteinů (viz obrázky 3-12, 3-13 a 3-14. Čísla v pravých dolních rozích jsou korelační koeficienty.).




*Obrázek 3-12: Vztah rozpustnosti proteinů a jeho taxonomické vzdálenosti od E. coli*



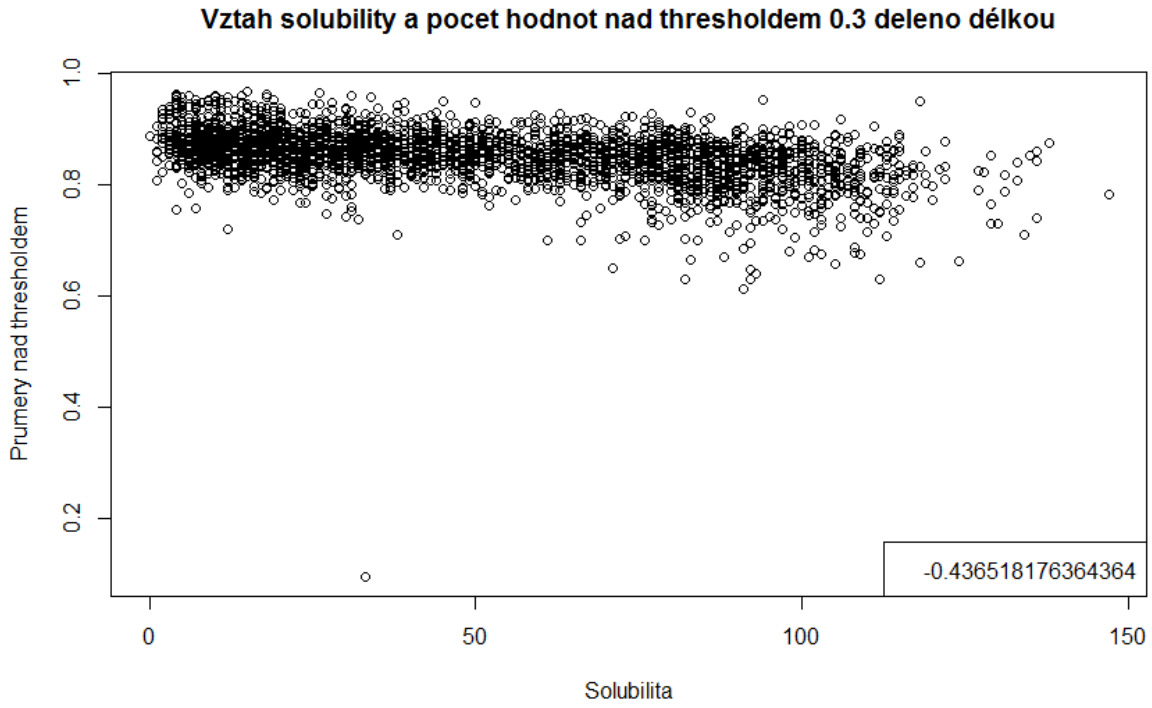

*Obrázek 3-13: Vztah rozpustnosti proteinů a jeho fylogenetické vzdálenosti od E. coli (počítáno z rRNA)*




*Obrázek 3-14: Vztah rozpustnosti proteinů a jeho fylogenetické vzdálenosti od E. coli (počítáno z ribozomů)*

### 3.2.4 Chaperony

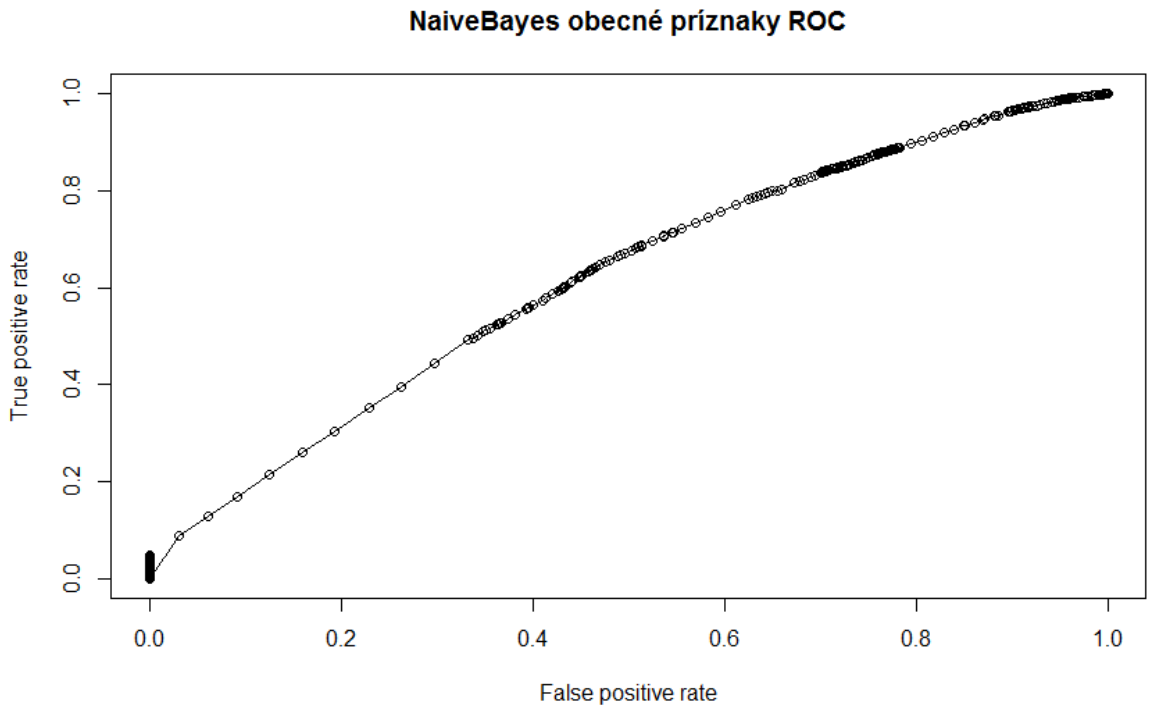
Jako potenciální příznaky jsme vybrali také potřebu proteinů využívat pomoci chaperonů ke správnému složení. Potřeba využití chaperonů byla určena pomoci nástroje BipPred (24), který predikuje místa na které se chaperony naváží. Tímto nástrojem byla otestována datová sada eSol. Výstupem nástroje BipPred je ale minimální a maximální profil nástroje. Z těchto dvou profilů byl vytvořen zprůměrováním průměrný profil a jako výstup byl vytvořen počet segmentů profilu, jejichž průměrná hodnota profilu přesahovala určitou mez. Meze byly testovány v intervalu 0.1–0.9 s krokem 0.1. Z grafů byla patrná podobnost s předchozími grafy zabývajícími se korelací délky a rozpustnosti na sadě eSol, a proto byla provedena korelace výsledků této metody s délkou. Byla zjištěná velice silná korelace, a proto byla celá metoda upravena. Počet segmentů nad mezí byl vydělen délkou a byl tak získán jakýsi procentuální poměr segmentů nad mezí vůči všem segmentům proteinu. Korelace výsledků nové metody s touto délkou prakticky zmizela od meze 0.3. Se zvyšující mezí se ale korelační koeficient snižoval, a tak byla mez 0.3 vybrána jako optimální. (viz obrázek 3-15). Veškeré grafy jsou dostupné na DVD.




*Obrázek 3-15: Vztah rozpustnosti proteinu a počtu hodnot nad mezí 0,3 děleném délkou proteinu na sadě eSol. Korelační koeficient v pravém dolním rohu.*

## 3.3 Návrh prediktoru

Jako příznaky pro tvorbu prediktoru byly vybrány taxonomická a fylogenetické vzdálenosti, doména zdrojového organismu, délka proteinu, GRAVY, náboj, aliphatický index, izoelektrický bod a počet sekvencí majících skóre BipPred nad mezemi 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 a 0.9. Tyto příznaky byly otestovány na sadě FitSet5 a výsledky zpracovány pomocí nástroje WEKA (25). Příznaky byly pomoci algoritmu *CfsSubsetEval* zredukovány za použití křížové validace s rozdělením na deset částí na doménu, náboj a izoelektrický bod. S těmito příznaky poté byly testovány *Classifier* algoritmy, opět za pomoci křížové validace s rozdělením na deset částí, a jako nejefektivnější se ukázal algoritmus *NaiveBayes* (ROC křivka viz obrázek 3-16).

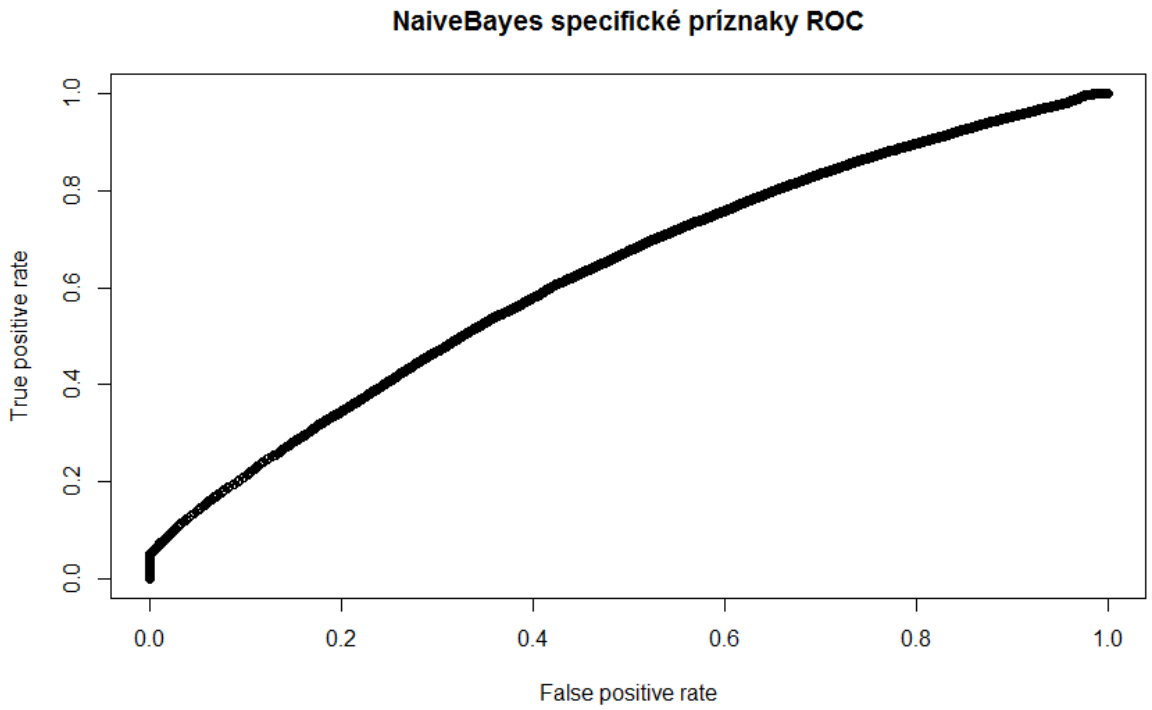



*Obrázek 3-16: NaiveBayes ROC křivka s použitím obecných příznaků, AUC = 0,6171,*

Originální příznaky byly poté opět zredukovány, tentokrát pomoci *ClassifierSubsetEval* speciálně po *NaiveBayes*. Výslednými příznaky byly doména, taxonomická vzdálenost, délka proteinu, GRAVY, náboj, izoelektrický bod a počet sekvencí majících skóre BipPred nad mezemi 0.2, 0.7, 0.8 a 0.9 (viz tabulka 1). Toto vedlo k velice mírném zlepšení predikce (ROC křivka viz obrázek 3-17).

| | Doména | Taxonomická vzdálenost | Vlastnosti proteinu | BipPred výsledky | Všechny příznaky |
|---|---|---|---|---|---|
| AUC ROC | 0,56 | 0,54 | 0,58 | 0,56 | 0,62 |
| poměr FP | 0,54 | 0,53 | 0,44 | 0,45 | 0,42 |
| poměr TP | 0,59 | 0,59 | 0,55 | 0,53 | 0,59 |

*Tabulka 1: Výsledky testování vybraných příznaků*




*Obrázek 3-17: NaiveBayes ROC křivka s použitím specifických příznaků, AUC = 0,6264*

# 4 Závěr

Tato práce navazovala na práci zabývající se predikcí agregace proteinů. Očekávali jsme že agregace proteinů a rozpustnost proteinů budou velice úzce spojeny, což platí alespoň z části u menší sady eSol, ale nikoliv u větších sad, založených na Target Tracku.

Bohužel se nepodařilo vytvořit kvalitní prediktor rozpustnosti proteinů. Bylo ale otestováno mnoho příznaků používaných ve starších nástrojích. Zjistili jsme, že na větší datové sadě jsou korelace jednotlivých příznaků a rozpustnosti proteinů menší, a i když například potřeba chaperonů vypadala při testování na datové sadě eSol nadějně, při testování na FitSet5 nevykazovala zdaleka tak dobré výsledky.

Jednou z možných cest pro dosažení větší přesnosti prediktoru je další rozšiřování a upravování datové sady používané při trénování a testování nástroje. Druhou možností je rozšiřování množiny vhodných příznaků. Ať už testováním příznaků použitých ve starších nástrojích, které nebyly na nové datové sadě otestovány. Například hledání specifických sekvencí aminokyselin v proteinech, jak to dělá například nástroj PROSO. Anebo rozšířením vstupů nástroje a přidáním informací o struktuře, pH atd. do vstupu nástroje.

# Literatura


1. NIWA, Tatsuya, Bei-Wen YING, Katsuyo SAITO, WenZhen JIN, Shoji TAKADA, Takuya UEDA a Hideki TAGUCHI. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences* [online]. 2009, **106**(11), 4201–4206. ISSN 0027-8424, 1091-6490. Dostupné z: doi:10.1073/pnas.0811922106

2. BAIROCH, Amos a Rolf APWEILER. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*. 2000, **28**(1), 45–48. ISSN 0305-1048.

3. *RCSB Protein Data Bank - RCSB PDB* [online]. [vid. 2017-05-09]. Dostupné z: http://www.rcsb.org/pdb/home/home.do

4. *TargetTrack* [online]. [vid. 2017-05-09]. Dostupné z: http://sbkb.org/tt/

5. SHIMIZU, Y., A. INOUE, Y. TOMARI, T. SUZUKI, T. YOKOGAWA, K. NISHIKAWA a T. UEDA. Cell-free translation reconstituted with purified components. *Nature Biotechnology* [online]. 2001, **19**(8), 751–755. ISSN 1087-0156. Dostupné z: doi:10.1038/90802

6. WILKINSON, D. L. a R. G. HARRISON. Predicting the solubility of recombinant proteins in Escherichia coli. *Bio/Technology (Nature Publishing Company)*. 1991, **9**(5), 443–448. ISSN 0733-222X.

7. IDICULA-THOMAS, Susan, Abhijit J. KULKARNI, Bhaskar D. KULKARNI, Valadi K. JAYARAMAN a Petety V. BALAJI. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics (Oxford, England)* [online]. 2006, **22**(3), 278–284. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/bti810

8. SMIALOWSKI, Pawel, Antonio J. MARTIN-GALIANO, Aleksandra MIKOLAJKA, Tobias GIRSCHICK, Tad A. HOLAK a Dmitrij FRISHMAN. Protein solubility: sequence based prediction and experimental verification. *Bioinformatics (Oxford, England)* [online]. 2007, **23**(19), 2536–2542. ISSN 1367-4811. Dostupné z: doi:10.1093/bioinformatics/btl623

9. MAGNAN, Christophe N., Arlo RANDALL a Pierre BALDI. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics (Oxford, England)* [online]. 2009, **25**(17), 2200–2207. ISSN 1367-4811. Dostupné z: doi:10.1093/bioinformatics/btp386

10. HUANG, Hui-Ling, Phasit CHAROENKWAN, Te-Fen KAO, Hua-Chin LEE, Fang-Lin CHANG, Wen-Lin HUANG, Shinn-Jang HO, Li-Sun SHU, Wen-Liang CHEN a Shinn-Ying HO. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC bioinformatics* [online]. 2012, **13 Suppl 17**, S3. ISSN 1471-2105. Dostupné z: doi:10.1186/1471-2105-13-S17-S3

11. SMIALOWSKI, Pawel, Gero DOOSE, Phillipp TORKLER, Stefanie KAUFMANN a Dmitrij FRISHMAN. PROSO II--a new method for protein solubility prediction. *The FEBS journal* [online]. 2012, **279**(12), 2192–2200. ISSN 1742-4658. Dostupné z: doi:10.1111/j.1742-4658.2012.08603.x

12. HIROSE, Shuichi a Tamotsu NOGUCHI. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics* [online]. 2013, **13**(9), 1444–1456. ISSN 1615-9861. Dostupné z: doi:10.1002/pmic.201200175

13. AGOSTINI, Federico, Michele VENDRUSCOLO a Gian Gaetano TARTAGLIA. Sequence-based prediction of protein solubility. *Journal of Molecular Biology* [online]. 2012, **421**(2-3), 237–241. ISSN 1089-8638. Dostupné z: doi:10.1016/j.jmb.2011.12.005

14. AGOSTINI, Federico, Davide CIRILLO, Carmen Maria LIVI, Riccardo DELLI PONTI a Gian Gaetano TARTAGLIA. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics (Oxford, England)* [online]. 2014, **30**(20), 2975–2977. ISSN 1367-4811. Dostupné z: doi:10.1093/bioinformatics/btu420

15. SORMANNI, Pietro, Francesco A. APRILE a Michele VENDRUSCOLO. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology* [online]. 2015, **427**(2), 478–490. ISSN 0022-2836. Dostupné z: doi:10.1016/j.jmb.2014.09.026

16. DE BAETS, Greet, Joost VAN DURME, Rob VAN DER KANT, Joost SCHYMKOWITZ a Frederic ROUSSEAU. Solubis: optimize your protein. *Bioinformatics (Oxford, England)* [online]. 2015, **31**(15), 2580–2582. ISSN 1367-4811. Dostupné z: doi:10.1093/bioinformatics/btv162

17. CHEN, Li, Rose OUGHTRED, Helen M. BERMAN a John WESTBROOK. TargetDB: a target registration database for structural genomics projects. *Bioinformatics (Oxford, England)* [online]. 2004, **20**(16), 2860–2862. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/bth300

18. KROGH, A., B. LARSSON, G. VON HEIJNE a E. L. SONNHAMMER. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* [online]. 2001, **305**(3), 567–580. ISSN 0022-2836. Dostupné z: doi:10.1006/jmbi.2000.4315

19. FU, Limin, Beifang NIU, Zhengwei ZHU, Sitao WU a Weizhong LI. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* [online]. 2012, **28**(23), 3150–3152. ISSN 1367-4811. Dostupné z: doi:10.1093/bioinformatics/bts565

20. FERNANDEZ-ESCAMILLA, Ana-Maria, Frederic ROUSSEAU, Joost SCHYMKOWITZ a Luis SERRANO. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* [online]. 2004, **22**(10), 1302–1306. ISSN 1087-0156. Dostupné z: doi:10.1038/nbt1012

21. ČERMÁK JIŘÍ. *Predikce vlivu aminokyselinových mutací na agregaci proteinu*. Brno, 2016. technická zpráva. FIT VUT v Brně.

22. TARTAGLIA, Gian Gaetano a Michele VENDRUSCOLO. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews* [online]. 2008, **37**(7), 1395–1401. ISSN 0306-0012. Dostupné z: doi:10.1039/b706784b

23. MAURER-STROH, Sebastian, Maja DEBULPAEP, Nico KUEMMERER, Manuela LOPEZ DE LA PAZ, Ivo Cristiano MARTINS, Joke REUMERS, Kyle L. MORRIS, Alastair COPLAND, Louise SERPELL, Luis SERRANO, Joost W. H. SCHYMKOWITZ a Frederic ROUSSEAU. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods* [online]. 2010, **7**(3), 237–242. ISSN 1548-7105. Dostupné z: doi:10.1038/nmeth.1432

24. SCHNEIDER, Markus, Mathias ROSAM, Manuel GLASER, Atanas PATRONOV, Harpreet SHAH, Katrin Christiane BACK, Marina Angelika DAAKE, Johannes BUCHNER a Iris ANTES. BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. *Proteins: Structure, Function, and Bioinformatics* [online]. 2016, **84**(10), 1390–1407. ISSN 1097-0134. Dostupné z: doi:10.1002/prot.25084

25. HALL, Mark, Eibe FRANK, Geoffrey HOLMES, Bernhard PFAHRINGER, Peter REUTEMANN a Ian H. WITTEN. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* [online]. 2009, **11**(1), 10–18. ISSN 1931-0145. Dostupné z: doi:10.1145/1656274.1656278

# Příloha 1. Obsah DVD

Přiložené DVD obsahuje tyto složky:

**BipPred**

Tato složka obsahuje dodatečné grafy s výsledky nástroje BipPred a dva skripty:

- run_bippred.php – skript, který automaticky spouští BipPred a ukládá výsledky
- TargetTrackByOneRun.py – skript, který dává předchozím skriptu data po jednom řádku.

**Tango APR**

Tato složka opět obsahuje dodatečné grafy s výsledky nástroje Tango

**eSol**

Složka eSol obsahuje skript Compare_Data.R, který zpracovává výsledky nástroje BipPred a vytváří z nich grafy

**Target Track**

Složka obsahuje tyto skripty:

- TargetTrackDatabaseToArff.py převádí soubory z csv do arff formátu
- SortTargetTrackUnsorted.py shlukuje databázi a vyřazuje z ní membránové proteiny. Tento skript pracuje s nástrojem CD-HIT a TMHMM a vyžaduje je ke své správné funkčnosti.
- MergeSortedAndPullThemFromOutput.py vyrovná shlukované sady a z originálního TargetTrack.csv souboru vybere korespondující záznamy.
- 0Compare_Data.R – podobně jako u eSol zpracovává výsledky nástroje BipPred

**Vlastnosti**

Složka obsahuje dva skripty:

- Features.py který obsahuje algoritmy pro výpočet jednotlivých vlastností
- MakeArffSet.py který vypočte vlastnost a uloží výsledek do arff formátu. Tento skript čte ze souboru database.txt, kde je potřeba data vložit ve formátu sekvence {mezera} id sekvence {konec řádku}

Dále DVD obsahuje skript MergeSets.R, který spojuje několik souborů arff do jednoho souboru pro hladké zpracování nástrojem Weka a zdrojový kód této práce.